Random sampling locations for comparing a mean with a fixed threshold (parametric)

Summary
This report summarizes the sampling design, associated statistical assumptions, as well as general guidelines for conducting post-sampling data analysis. Sampling plan components presented here include how many sampling locations to choose and where within the sampling area to collect those samples. The type of medium to sample (i.e., soil, groundwater, etc.) and how to analyze the samples (in-situ, fixed laboratory, etc.) are addressed in other sections of the sampling plan.

The following table summarizes the sampling design. A figure that shows sampling locations in the field and a table that lists sampling location coordinates are also provided below.

SUMMARY OF SAMPLING DESIGN	
Primary Objective of Design	Compare a site mean to a fixed threshold
Type of Sampling Design	Parametric
Sample Placement (Location) in the Field	Simple random sampling
Working (Null) Hypothesis	The mean value at the site exceeds the threshold
Formula for calculating number of sampling locations	Student's t-test
Calculated total number of samples	79
Number of samples on map [a]	79
Number of selected sample areas [b]	2
Specified sampling area [c]	188054.34 m^2
Total cost of sampling [d]	$40,500.00

[a] This number may differ from the calculated number because of 1) grid edge effects, 2) adding judgment samples, or 3) selecting or unselecting sample areas.
[b] The number of selected sample areas is the number of colored areas on the map of the site. These sample areas contain the locations where samples are collected.
[c] The sampling area is the total surface area of the selected colored sample areas on the map of the site.
[d] Including measurement analyses and fixed overhead costs. See the Cost of Sampling section for an explanation of the costs presented here.
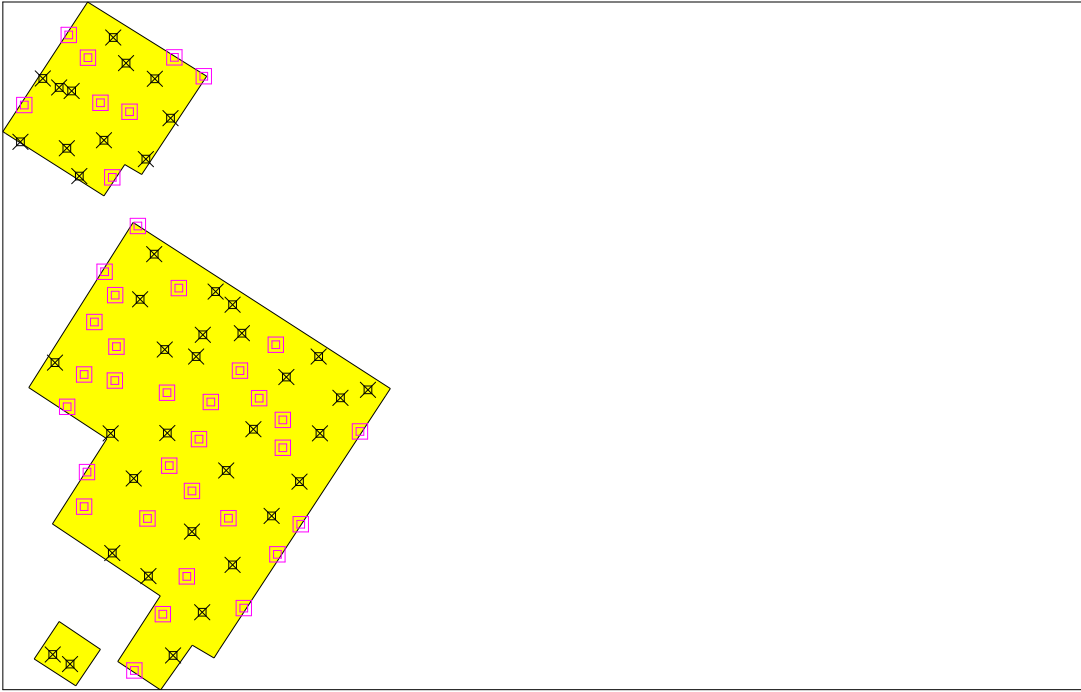
Area: Area 1					
X Coord	**Y Coord**	**Label**	**Value**	**Type**	**Historical**
679242.7260	3083326.5280	J-07S	0.47	Manual	T
679104.2450	3083223.2620	J-02S	0.34	Manual	T
679181.2750	3083178.2880	J-08S	0.29	Manual	T
679133.4290	3083306.3130	J-01S	0.14	Manual	T
679268.7700	3083200.3260	J-11S	0.26	Manual	T
679171.2970	3083289.7960	J-04S	0.115	Manual	T
679164.8060	3083214.7100	J-06S	0.23	Manual	T
679213.7730	3083224.9730	J-09S	0.105	Manual	T
679280.5440	3083305.6810	J-10S	0.105	Manual	T
679155.0740	3083294.6960	J-03S	0.1	Manual	T
679301.1600	3083254.0340	J-12S	0.09	Manual	T
679225.8560	3083359.9740	J-05S	0.08	Manual	T
679344.9917	3083309.2278	J-38S	2.2	Adaptive-Fill	
679167.5023	3083363.3712	J-37S	2.1	Adaptive-Fill	
679247.5039	3083262.4916	J-19S	1.7	Adaptive-Fill	
679108.9961	3083271.3916	J-36S	1.7	Adaptive-Fill	
679224.6687	3083176.2417	J-41S	1.7	Adaptive-Fill	
679192.8634	3083333.4953	J-23S	1.4	Adaptive-Fill	
679208.9664	3083274.1678	J-28S	1.2	Adaptive-Fill	
679306.1204	3083334.0306	J-32S	1.1	Adaptive-Fill	

Area: Area 3

X Coord	Y Coord	Label	Value	Type	Historical
679470.3570	3082776.7350	J-38S	2.2	Manual	T
679433.9450	3082731.6820	J-37S	2.1	Manual	T
679343.5810	3082969.5980	J-19S	1.7	Manual	T
679382.8900	3082667.5270	J-36S	1.7	Manual	T
679560.6070	3082897.2580	J-41S	1.7	Manual	T
679297.0010	3082840.6970	J-23S	1.4	Manual	T
679272.0040	3082652.6750	J-28S	1.2	Manual	T
679453.4760	3082914.1150	J-32S	1.1	Manual	T
679146.6460	3082549.7640	J-25S	1	Manual	T
679169.0760	3082537.3510	J-27S	1	Manual	T
679495.8840	3082940.9730	J-33S	0.81	Manual	T
679304.6530	3082548.6880	J-34S	0.8	Manual	T
679342.7410	3082605.3190	J-35S	0.78	Manual	T
679252.7130	3082781.0290	J-22S	0.74	Manual	T
679497.3310	3082840.3960	J-39S	0.7	Manual	T
679524.3310	3082886.8990	J-40S	0.62	Manual	T
679329.4380	3082711.0960	J-29S	0.58	Manual	T
679149.4920	3082933.0980	J-13S	0.54	Manual	T
679374.4420	3082791.3300	J-30S	0.54	Manual	T
679360.5700	3083026.4980	J-18S	0.44	Manual	T
679293.5600	3082950.4980	J-17S	0.41	Manual	T
679261.0980	3083016.3510	J-15S	0.38	Manual	T
679222.6340	3082840.1720	J-16S	0.33	Manual	T
679394.8070	3082971.8300	J-24S	0.3	Manual	T
679382.8640	3083009.1130	J-20S	0.12	Manual	T
679279.6830	3083075.4290	J-14S	0.115	Manual	T
679410.1490	3082845.8460	J-31S	0.23	Manual	T
679224.5850	3082683.1400	J-26S	0.11	Manual	T
679335.0020	3082941.1720	J-21S	0.1	Manual	T
679187.7108	3082743.9049		0	Adaptive-Fill	
679201.3419	3082986.3558		0	Adaptive-Fill	
679228.1261	3082909.4864		0	Adaptive-Fill	
679354.1234	3082881.5703		0	Adaptive-Fill	
679165.3089	3082875.0841		0	Adaptive-Fill	
679214.7953	3083052.5423		0	Adaptive-Fill	
679270.7295	3082728.2056		0	Adaptive-Fill	
679253.6530	3082528.8006		0	Adaptive-Fill	
679296.7054	3082893.3847		0	Adaptive-Fill	
679329.1669	3082764.4865		0	Adaptive-Fill	
679322.5779	3082652.2311		0	Adaptive-Fill	

679377.2589	3082728.9013	0	Adaptive-Fill
679397.3161	3082610.7954	0	Adaptive-Fill
679392.4419	3082922.3762	0	Adaptive-Fill
679291.0689	3082602.5607	0	Adaptive-Fill
679191.3863	3082789.3978	0	Adaptive-Fill
679439.5764	3082956.5943	0	Adaptive-Fill
679441.5982	3082681.4673	0	Adaptive-Fill
679448.5842	3082821.4334	0	Adaptive-Fill
679299.4420	3082797.6478	0	Adaptive-Fill
679549.8870	3082842.9025	0	Adaptive-Fill
679312.1684	3083030.8840	0	Adaptive-Fill
679230.0779	3082953.9913	0	Adaptive-Fill
679338.6362	3082832.5643	0	Adaptive-Fill
679472.1177	3082720.7150	0	Adaptive-Fill
679258.2576	3083112.5898	0	Adaptive-Fill
679417.6739	3082886.3804	0	Adaptive-Fill
679187.6340	3082917.6887	0	Adaptive-Fill
679448.6706	3082857.7165	0	Adaptive-Fill
679228.4704	3083021.7000	0	Adaptive-Fill

Primary Sampling Objective
The primary purpose of sampling at this site is to compare a mean value of a site with a fixed threshold. The working hypothesis (or 'null' hypothesis) is that the mean value at the site is equal to or exceeds the threshold. The alternative hypothesis is that the mean value is less than the threshold. VSP calculates the number of samples required to reject the null hypothesis in favor of the alternative hypothesis, given a selected sampling approach and inputs to the associated equation.

Selected Sampling Approach
A parametric random sampling approach was used to determine the number of samples and to specify sampling locations. A parametric formula was chosen because the conceptual model and historical information (e.g., historical data from this site or a very similar site) indicate that parametric assumptions are reasonable. These assumptions will be examined in post-sampling data analysis.

Both parametric and non-parametric approaches rely on assumptions about the population. However, non-parametric approaches typically require fewer assumptions and allow for more uncertainty about the statistical distribution of values at the site. The trade-off is that if the parametric assumptions are valid, the required number of samples is usually less than the number of samples required by non-parametric approaches.

Locating the sample points randomly provides data that are separated by many distances, whereas systematic samples are all equidistant apart. Therefore, random sampling provides more information about the spatial structure of the potential contamination than systematic sampling does. As with systematic sampling, random sampling also provides information regarding the mean value, but there is the possibility that areas of the site will not be represented with the same frequency as if uniform grid sampling were performed.

Number of Total Samples: Calculation Equation and Inputs
The equation used to calculate the number of samples is based on a Student's t-test. For this site, the null hypothesis is rejected in favor of the alternative hypothesis if the sample mean is sufficiently smaller than the threshold. The number of samples to collect is calculated so that 1) there will be a high probability $(1-\beta)$ of rejecting the null hypothesis if the alternative hypothesis is true and 2) a low probability (α) of rejecting the null hypothesis if the null hypothesis is true.

The formula used to calculate the number of samples is:

$$n = \frac{S^2}{\Delta^2}\left(Z_{1-\alpha} + Z_{1-\beta}\right)^2 + 0.5Z_{1-\alpha}^2$$

where

n is the number of samples,
S is the estimated standard deviation of the measured values including analytical error,
Δ is the width of the gray region,
α is the acceptable probability of incorrectly concluding the site mean is less than the threshold,
β is the acceptable probability of incorrectly concluding the site mean exceeds the threshold,
$Z_{1-\alpha}$ is the value of the standard normal distribution such that the proportion of the distribution less than $Z_{1-\alpha}$ is 1-α,
$Z_{1-\beta}$ is the value of the standard normal distribution such that the proportion of the distribution less than $Z_{1-\beta}$ is 1-β.

The values of these inputs that result in the calculated number of sampling locations are:

| Analyte	n	Parameter					
		S	Δ	α	β	$Z_{1-\alpha}$ [a]	$Z_{1-\beta}$ [b]
Arsenic	79	0.58368 mg/kg	0.1948 mg/kg	0.05	0.1	1.64485	1.28155

[a] This value is automatically calculated by VSP based upon the user defined value of α.
[b] This value is automatically calculated by VSP based upon the user defined value of β.

The following figure is a performance goal diagram, described in EPA's QA/G-4 guidance (EPA, 2000). It shows the probability of concluding the sample area is dirty on the vertical axis versus a range of possible true mean values for the site on the horizontal axis. This graph contains all of the inputs to the number of samples equation and pictorially represents the calculation.

The red vertical line is shown at the threshold (action limit) on the horizontal axis. The width of the gray shaded area is equal to Δ; the upper horizontal dashed blue line is positioned at 1-α on the vertical axis; the lower horizontal dashed blue line is positioned at β on the vertical axis. The vertical green line is positioned at one standard deviation below the threshold. The shape of the red curve corresponds to the estimates of variability. The calculated number of samples results in the curve that passes through the lower bound of Δ at β and the upper bound of Δ at 1-α. If any of the inputs change, the number of samples that result in the correct curve changes.

1-Sample t-Test of True Mean vs. Action Level

n=79, alpha=5%, beta=10%, std.dev.=0.58368
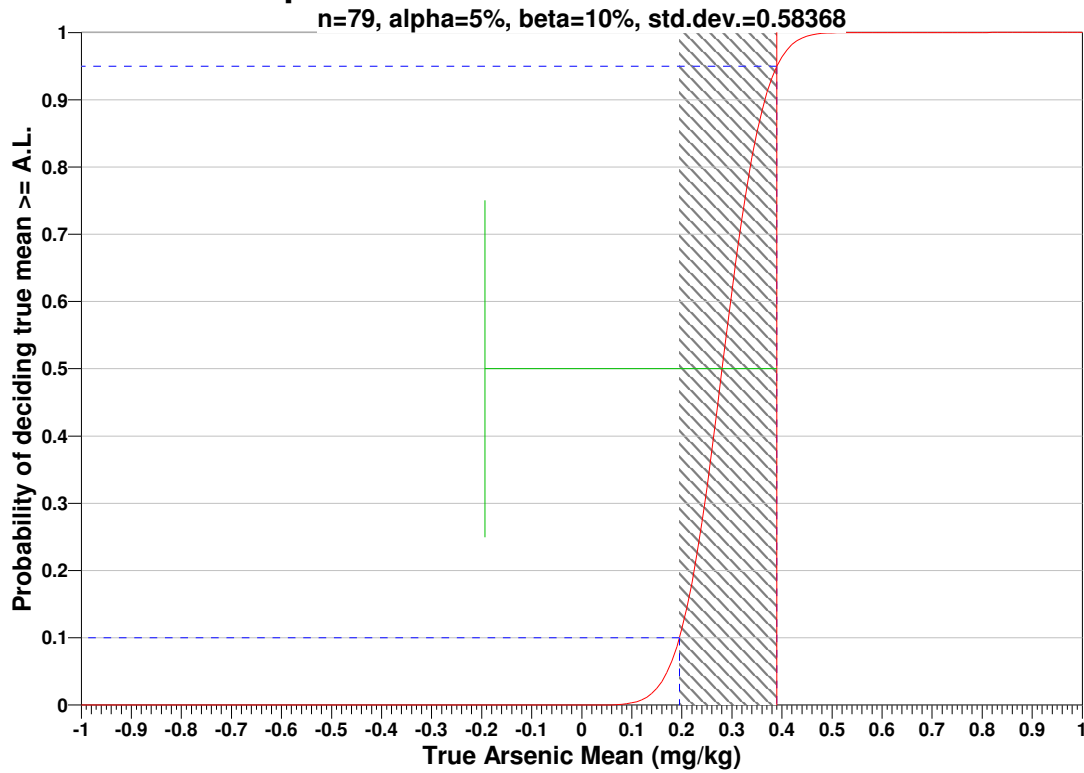
Statistical Assumptions

The assumptions associated with the formulas for computing the number of samples are:
1.   the sample mean is normally distributed (this happens if the data are roughly symmetric or the sample size is more than 30; for extremely skewed data sets, additional samples may be required for the sample mean to be normally distributed),
2.   the variance estimate, S^2, is reasonable and representative of the population being sampled,
3.   the population values are not spatially or temporally correlated, and
4.   the sampling locations will be selected randomly.

The first three assumptions will be assessed in a post data collection analysis. The last assumption is valid because the sample locations were selected using a random process.

Sensitivity Analysis

The sensitivity of the calculation of number of samples was explored by varying the standard deviation, lower bound of gray region (% of action level), beta (%), probability of mistakenly concluding that μ > action level and alpha (%), probability of mistakenly concluding that μ < action level. The following table shows the results of this analysis.

Number of Samples							
AL=0.39		**α=5**		**α=10**		**α=15**	
		s=1.16736	s=0.58368	s=1.16736	s=0.58368	s=1.16736	s=0.58368
LBGR=90	**β=5**	9698	2426	7674	1920	6442	1611
	β=10	7675	1920	5887	1473	4815	1205
	β=15	6443	1612	4815	1205	3851	963
LBGR=80	**β=5**	2426	608	1920	481	1611	404
	β=10	1920	481	1473	369	1205	302
	β=15	1612	404	1205	302	963	242
LBGR=70	**β=5**	1079	271	854	214	717	180

β=10	854	215	655	165	536	135
β=15	718	181	536	135	429	108

s = Standard Deviation
LBGR = Lower Bound of Gray Region (% of Action Level)
β = Beta (%), Probability of mistakenly concluding that μ > action level
α = Alpha (%), Probability of mistakenly concluding that μ < action level
AL = Action Level (Threshold)

Cost of Sampling
The total cost of the completed sampling program depends on several cost inputs, some of which are fixed, and others that are based on the number of samples collected and measured. Based on the numbers of samples determined above, the estimated total cost of sampling and analysis at this site is $40,500.00, which averages out to a per sample cost of $512.66. The following table summarizes the inputs and resulting cost estimates.

COST INFORMATION			
Cost Details	Per Analysis	Per Sample	79 Samples
Field collection costs		$100.00	$7,900.00
Analytical costs	$400.00	$400.00	$31,600.00
Sum of Field & Analytical costs		**$500.00**	**$39,500.00**
Fixed planning and validation costs			$1,000.00
Total cost			**$40,500.00**

Data Analysis for Arsenic
The following data points were entered by the user for analysis.

Arsenic (mg/kg)										
Rank	1	2	3	4	5	6	7	8	9	10
0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0
30	0.08	0.09	0.1	0.1	0.105	0.105	0.11	0.115	0.115	0.12
40	0.14	0.23	0.23	0.26	0.29	0.3	0.33	0.34	0.38	0.41
50	0.44	0.47	0.54	0.54	0.58	0.62	0.7	0.74	0.78	0.8
60	0.81	1	1	1.1	1.1	1.2	1.2	1.4	1.4	1.7
70	1.7	1.7	1.7	1.7	1.7	2.1	2.1	2.2	2.2	

SUMMARY STATISTICS for Arsenic	
n	79
Min	0
Max	2.2
Range	2.2
Mean	0.49582
Median	0.12
Variance	0.42709

StdDev	0.65352
Std Error	0.073527
Skewness	1.2833
Interquartile Range	0.8

Percentiles								
1%	5%	10%	25%	50%	75%	90%	95%	99%
0	0	0	0	0.12	0.8	1.7	2.1	2.2

Outlier Test

Rosner's test for multiple outliers was performed to test whether the most extreme value is a statistical outlier. The test was conducted at the 5% significance level.

Data should not be excluded from analysis solely on the basis of the results of this or any other statistical test. If any values are flagged as possible outliers, further investigation is recommended to determine whether there is a plausible explanation that justifies removing or replacing them.

In using Rosner's test to detect up to 1 outlier, a test statistic R_1 is calculated, and compared with a critical value C_1 to test the hypothesis that there is one outlier in the data.

ROSNER'S OUTLIER TEST for Arsenic			
k	Test Statistic R_k	5% Critical Value C_k	Significant?
1	2.68	3.05	No

None of the test statistics exceeded the corresponding critical values, therefore none of the 1 tests are significant and we conclude that at the 5% significance level there are no outliers in the data.

A normal distribution test indicated that the data do not appear to be normally distributed, so further investigation is recommended before using the results of this test. Because Rosner's test can be used only when the data without the suspected outlier are approximately normally distributed, a Lilliefors test for normality was performed at a 5% significance level.

NORMAL DISTRIBUTION TEST (excluding outliers)	
Lilliefors Test Statistic	0.8461
Lilliefors 5% Critical Value	0.94

The calculated Lilliefors test statistic is less than the 5% Lilliefors critical value, so the test rejects the hypothesis that the data are normal and concludes that the data, excluding the most extreme value, do not appear to follow a normal distribution at the 5% level of significance. Rosner's test may not be appropriate if the assumption of normally distributed data is not justified for this data set. Examine the Q-Q plot displayed below to further assess the normality of the data.

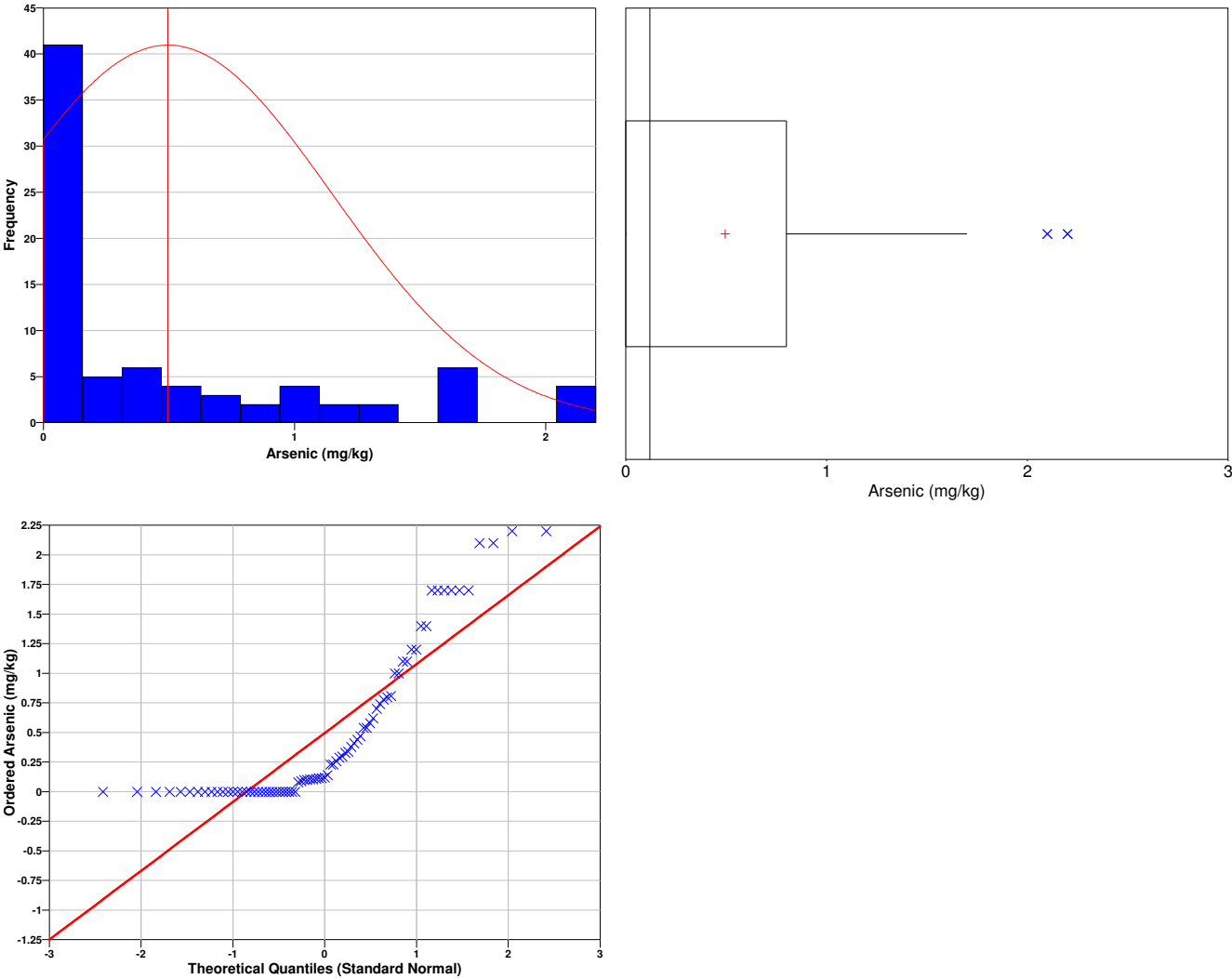**Data Plots for Arsenic**

Graphical displays of the data are shown below.

The Histogram is a plot of the fraction of the n observed data that fall within specified data "bins." A histogram is generated by dividing the x axis (range of the observed data values) into "bins" and displaying the number of data in each bin as the height of a bar for the bin. The area of the bar is the fraction of the n data values that lie within the bin. The sum of the fractions for all bins equals one. A histogram is used to assess how the n data are distributed (spread) over their range of values. If the histogram is more or less symmetric and bell shaped, then the data may be normally distributed.

The Box and Whiskers plot is composed of a central box divided by a line, and with two lines extending out from the box, called the "whiskers". The line through the box is drawn at the median of the n data observed. The two ends of the box represent the 25th and 75th percentiles of the n data values, which are also called the lower and upper quartiles, respectively, of the data set. The sample mean (mean of the n data) is shown as a "+" sign. The upper whisker extends

to the largest data value that is less than the upper quartile plus 1.5 times the interquartile range (upper quartile minus the lower quartile). The lower whisker extends to the smallest data value that is greater than the lower quartile minus 1.5 times the interquartile range. Extreme data values (greater or smaller than the ends of the whiskers) are plotted individually as blue Xs. A Box and Whiskers plot is used to assess the symmetry of the distribution of the data set. If the distribution is symmetrical, the box is divided into two equal halves by the median, the whiskers will be the same length, and the number of extreme data points will be distributed equally on either end of the plot.

The Q-Q plot graphs the quantiles of a set of n data against the quantiles of a specific distribution. We show here only the Q-Q plot for an assumed normal distribution. The pth quantile of a distribution of data is the data value, x_n, for which a fraction p of the distribution is less than x_n. If the data plotted on the normal distribution Q-Q plot closely follow a straight line, even at the ends of the line, then the data may be assumed to be normally distributed. If the data points deviate substantially from a linear line, then the data are not normally distributed.

For more information on these plots consult Guidance for Data Quality Assessment, EPA QA/G-9, pgs 2.3-1 through 2.3-12. (http://www.epa.gov/quality/qa-docs.html).

Tests for Arsenic
A goodness-of-fit test was performed to test whether the data set had been drawn from an underlying normal distribution. The Lilliefors test was used to test the null hypothesis that the data are normally distributed. The test was conducted at the 5% significance level, i.e., the probability the test incorrectly rejects the null hypothesis was set at 0.05.

NORMAL DISTRIBUTION TEST	
Lilliefors Test Statistic	0.2259
Lilliefors 5% Critical Value	0.09968

The calculated Lilliefors test statistic exceeds the 5% Lilliefors critical value, so we can reject the hypothesis that the data are normal, or in other words the data do not appear to follow a normal distribution at the 5% level of significance. The Q-Q plot displayed above should be used to further assess the normality of the data.

Upper Confidence Limit on the True Mean
Two methods were used to compute the upper confidence limit (UCL) on the mean. The first is a parametric method that assumes a normal distribution. The second is the Chebyshev method, which requires no distributional assumption.

UCLs ON THE MEAN	
95% Parametric UCL	0.6182
95% Non-Parametric (Chebyshev) UCL	0.8163

Because the data do not appear to be normally distributed according to the goodness-of-fit test performed above, the non-parametric UCL (0.8163) may be a more accurate upper confidence limit on the true mean.

One-Sample t-Test
A one-sample t-test was performed to compare the sample mean to the action level. The null hypothesis used is that the true mean equals or exceeds the action level (AL). The t-test was conducted at the 5% significance level. The sample value t was computed using the following equation:

$$t = \frac{\bar{x} - AL}{SE}$$

where
x is the sample mean of the n=79 data,
AL is the action level or threshold (0.39),
SE is the standard error = (standard deviation) / (square root of n).

This t was then compared with the critical value $t_{0.95}$, where $t_{0.95}$ is the value of the t distribution with n-1=78 degrees of freedom for which the proportion of the distribution to the left of $t_{0.95}$ is 0.95. The null hypothesis will be rejected if $t < -t_{0.95}$.

ONE-SAMPLE t-TEST		
t-statistic	Critical Value $t_{0.95}$	Null Hypothesis
1.4392	1.6646	Cannot Reject

The test did not reject the null hypothesis that the mean value at the site exceeds the threshold, therefore conclude the true mean exceeds the threshold.

Because the data do not appear to be normally distributed, the MARSSIM Sign Test might be preferred over the One Sample t-Test. The following table represents the results of the MARSSIM Sign Test using the current data:

MARSSIM Sign Test		
Test Statistic (S+)	95% Critical Value	Null Hypothesis
49	47	Reject